ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2017)
66, *Part* 4, *pp.* 869–890

Statistical inference of the mechanisms driving collective cell movement

Elaine A. Ferguson and Jason Matthiopoulos,

University of Glasgow, UK

Robert H. Insall

Cancer Research UK Beatson Institute, Glasgow, UK

and Dirk Husmeier

University of Glasgow, UK

[Received August 2015. Final revision October 2016]

Summary. Numerous biological processes, many impacting on human health, rely on collective cell movement. We develop nine candidate models, based on advection–diffusion partial differential equations, to describe various alternative mechanisms that may drive cell movement. The parameters of these models were inferred from one-dimensional projections of laboratory observations of *Dictyostelium discoideum* cells by sampling from the posterior distribution using the delayed rejection adaptive Metropolis algorithm. The best model was selected by using the widely applicable information criterion. We conclude that cell movement in our study system was driven both by a self-generated gradient in an attractant that the cells could deplete locally, and by chemical interactions between the cells.

Keywords: Advection–diffusion; Collective cell movement; Delayed rejection adaptive Metropolis algorithm; Model selection; Self-generated gradients; Widely applicable information criterion

1. Introduction

Collective movements of eukaryotic cells are essential for the occurrence of many major biological processes, such as tissue development, wound healing and cancer cell invasion and metastasis. The majority of the mechanisms that are proposed as drivers of cell movement invoke the process of chemotaxis, whereby cells bias their movement in response to gradients in the concentration of certain chemicals (chemoattractants) in their environment (Insall, 2010; Majumdar *et al*., 2014). This allows cells to track favourable conditions in spatiotemporally varying environments. The formation of chemotactic gradients may occur through the presence of a local chemoattractant source, e.g. macrophages releasing epidermal growth factor, which induces migration of breast tumour cells (Wyckoff *et al*., 2004). Alternatively, gradients can form through local depletion of a widely produced chemical. Local depletion has historically received less attention than local production as a gradient forming mechanism. However, several recent studies have revealed cases where cells move in response to gradients that they have created themselves by depletion of a chemoattractant, sparking new interest in the area of self-generated gradients

Address for correspondence: Elaine A. Ferguson, Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK.
E-mail: e.ferguson.2@research.gla.ac.uk

 0035–9254/17/66869
Published by John Wiley & Sons Ltd on behalf of the Royal Statistical Society.

(Scherber *et al*., 2012; Donà *et al*., 2013; Venkiteswaran *et al*., 2013; Muinonen-Martin *et al*., 2014; Tweedy *et al*., 2016).

Dictyostelium discoideum, which is an amoeba that can exist in both unicellular and multicellular forms, has emerged as a model organism for eukaryotic cell movement. The biological machinery underlying cell movement is very similar between *Dictyostelium* and the cells of higher animals, and movement of this organism can be driven through chemotaxis in exactly the same way as that of human leukocytes and invading cancer cells, for example, making it well suited as a vehicle for gaining insights on human disease (Carnell and Insall, 2011). When in their solitary form, *Dictyostelium* cells feed on bacteria, which are located by climbing up gradients in bacteria-produced chemicals. Solitary *Dictyostelium* cells also respond to chemical gradients under conditions of starvation, when they produce waves of the chemoattractant cyclic adenosine monophosphate, attracting other nearby cells and resulting in the formation of a multicellular aggregate that develops into a fruiting body, facilitating dispersal (Loomis, 1982). In this study, we took data on the movement of a group of *Dictyostelium* cells in the solitary phase and attempted to identify the mechanisms that are involved in producing this movement by using mathematical models of cell movement.

Distinguishing between competing models that describe movement in terms of alternative mechanisms increases our understanding of the movement process and how it might be manipulated. In cellular systems where movement has an effect on human health, such as cancer cell invasion and immune responses, these insights could be used in the development of medical interventions. For example, if a type of tumour cell moves in response to gradients in a certain chemoattractant, movement of this cell type could perhaps be managed by targeted release of the chemoattractant from an implant, either to restrict movement from the primary tumour or to redirect movement away from critical tissues (Fleming and Saltzman, 2002; Deisboeck and Couzin, 2009). Some cell interactions act to drive movement (Wyckoff *et al*., 2004), so medications that disrupt these communications could effectively limit movement, whereas other interactions inhibit movement (McDonough *et al*., 1999), so promoting them becomes the goal. Despite the importance of selecting between these different biological hypotheses for cell movement, attempts to fit cell movement models to data formally have been rare. Our objective, therefore, was to develop a methodology both for fitting a number of competing models to data on cell movement and for comparing these models on the basis of model selection criteria.

Although models that describe movement at the level of the individual cell (see, for example, Neilson *et al*. (2011) and Coburn *et al*. (2013)) can be effective for simulating movement patterns, inference on these models can be prohibitively slow because of high computational costs, making population-based approaches more attractive. A population-based cell movement model that has proved popular is the Keller–Segel model, which was developed to explore the aggregating behaviour of *Dictyostelium* (Keller and Segel, 1970, 1971). This model uses partial differential equations (PDEs) of the advection–diffusion type to describe how the distribution of cells in time and space is affected by a chemoattractant that is released by the cells into their environment and can also be depleted by the cells. Advection–diffusion equations describe movement of agents in terms of a directional component (advection) and a random component (diffusion). Since the introduction of the Keller–Segal model, a wide range of advection–diffusion models for cell movement, incorporating various cell behaviours, have been developed (see Hillen and Painter (2009) for a guide). Models of this type have also been widely used within the environmental and ecological literature, describing everything from the transport of pollutants in the atmosphere (Zlatev *et al*., 1984) to the movements of caribou in response to human disturbance (Fortin *et al*., 2013). There have been several efforts to carry out inference for advection–diffusion models in these fields. Maximum likelihood approaches have been applied to infer the parameters of

models for both tuna (Sibert *et al.*, 1999) and coyote packs (Moorcroft *et al.*, 2006), with model selection based on the Akaike information criterion also being carried out in the coyote case. Hierarchical Bayesian approaches to inference in these models have also been discussed and demonstrated on data on the invasion of North America by the Eurasian collared dove (Wikle and Hooten, 2006; Cressie and Wikle, 2011). We adopted this same flexible advection–diffusion modelling framework in this study. Although such models have previously been used to simulate cell movement, attempts to fit these models to data and to carry out model selection have been absent in this field.

Statistical inference for cell movement models poses several challenges. Analytical solutions of these models are typically unavailable, so we must resort to numerical solution; in the case of advection–diffusion equations, for example, analytical solutions are available only for certain functional forms of the advection and diffusion rates (Zoppou and Knight, 1997; Jaiswal and Kumar, 2011). Lack of a closed form likelihood can also necessitate numerical optimization, making inference computationally expensive. Inaccuracies in the numerical solution of PDE models result from the need to discretize time and space, so a balance between accuracy and model running costs must be established (Soetaert and Herman, 2009). Numerical solutions of advection–diffusion models have additional stability issues that can make fitting particularly challenging. The Péclet number is the ratio of the advection rate to the diffusion rate, multiplied by the grid spacing of the spatial discretization that is used to integrate the model (Soetaert and Herman, 2009). When this quantity exceeds 1 (i.e. advection is dominant), oscillating numerical solutions that would not occur analytically can be produced. This can cause failure of model solving algorithms, so certain regions of parameter space cannot be fully explored (Sibert *et al.*, 1999). Complex likelihood surfaces with many local optima can also make reaching the global optimum very challenging. Finally, adequate data on all important variables are not always available; cells may be affected by unidentified chemicals in their environment, and concentrations of even known important chemicals may be impossible to obtain at sufficiently high spatiotemporal resolution. In such cases, the fitting process is further complicated by the need to infer these latent variables from the information that is provided by the observed variables. Overcoming these difficulties in model fitting would be an important step towards helping us to understand cell movement in a wide range of systems.

Here, we introduce a set of competing models using an advection–diffusion framework to describe cell movement. The movement mechanisms that are incorporated in these models include cell responses to both a gradient in a chemoattractant that can be broken down by the cells and a gradient in conspecific density. Using data on *Dictyostelium* movement, we carry out inference for these models by sampling parameters from the posterior distribution using the delayed rejection adaptive Metropolis algorithm DRAM (Haario *et al.*, 2006). Model comparison was then carried out by using the widely applicable information criterion (WAIC) (Watanabe, 2010).

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Data

The two data sets that are utilized in our analyses were collected by Tweedy *et al.* (2016) using an assay designed to observe the movement of a group of *Dictyostelium* amoebae under agarose (Laevsky and Knecht, 2001) in response to a self-generated gradient in the chemoattractant folate, which can be broken down by the cells. *Dictyostelium* cells were added to a trough cut

into a dish of agarose, which in one case contained homogeneously distributed folate at a concentration of 10 μM, and in the second case contained no folate. In both cases, there was no folate in the trough containing the cells. The cells were given an hour to adhere to the bottom of the dish and the movement of the cells across one edge of the trough and under the agar was imaged under a microscope (Fig. 1(a)) over 5.5 h for the 10-μM folate data set and 3.5 h for the 0-μM folate data set. For both sets of images, the number of cells that were visible increased considerably over time, as more cells moved out of the high-density trough area and into the field of view.

The co-ordinates of the cells were manually extracted from the images at half-hourly intervals. Since all cells started the assay in the linear trough along the y-axis, and we were primarily interested in movement perpendicular to the trough, along the x-axis, the data set is effectively one dimensional (an additional analysis supporting this simplifying assumption is presented in the on-line supplement A). We, therefore, collapsed the data set along the y-axis for our analyses, considering only the x-co-ordinates of the cells. For the 10-μM folate data set, one-dimensional density estimates obtained from the cell location data show the gradual spread of the group of cells up the spatial axis, and reveal that, over time, the distribution of cells starts to become bimodal, with one peak indicating the progressing cell front and a second peak indicating the cells' point of origin at the edge of the trough (Figs 1(b)–1(m)). This peaked cell front is not visible in the 0-μM folate data, where cells move out from the trough more slowly and in lower densities (supplement B).

It was not possible to measure accurately the fine-scale spatiotemporal variation in folate concentration for the data set where this chemoattractant was present, so the changing distribution of this attractant could not be captured in the same way as the cell density. Therefore, attractant concentration was treated as a latent variable during the fitting of our cell movement models to these data.

3. Models

3.1. Model descriptions

The basic form of a one-dimensional advection–diffusion equation for cell movement is

$$\frac{\partial C(x,t)}{\partial t} = \underbrace{-\frac{\partial}{\partial x}\{a(x,t)\,C(x,t)\}}_{\text{advection}} + \underbrace{\frac{\partial}{\partial x}\left\{D_C(x,t)\,\frac{\partial C(x,t)}{\partial x}\right\}}_{\text{diffusion}} \tag{1}$$

where $C(x,t)$ is the cell density, $a(x,t)$ is the cell advection coefficient and $D_C(x,t)$ is the cell diffusion coefficient. The advection term implies directional movement up the spatial axis if $a(x,t)$ is positive and down the spatial axis if $a(x,t)$ is negative, with the speed of movement determined by the magnitude of $a(x,t)$. The diffusion term describes net movement of cells from regions of high to low density at a rate described by $D_C \geqslant 0$.

To investigate various alternative mechanisms for cell movement, we developed eight forms of the advection coefficient. In the *diffusion model*, movement occurs through diffusion only, i.e.

$$a(x,t) = 0. \tag{2}$$

In the *gradient model*, we introduced the effect of an attractant by assuming that directional cell movement occurs up a spatial gradient in attractant concentration $A(x,t)$, using the advection coefficient:

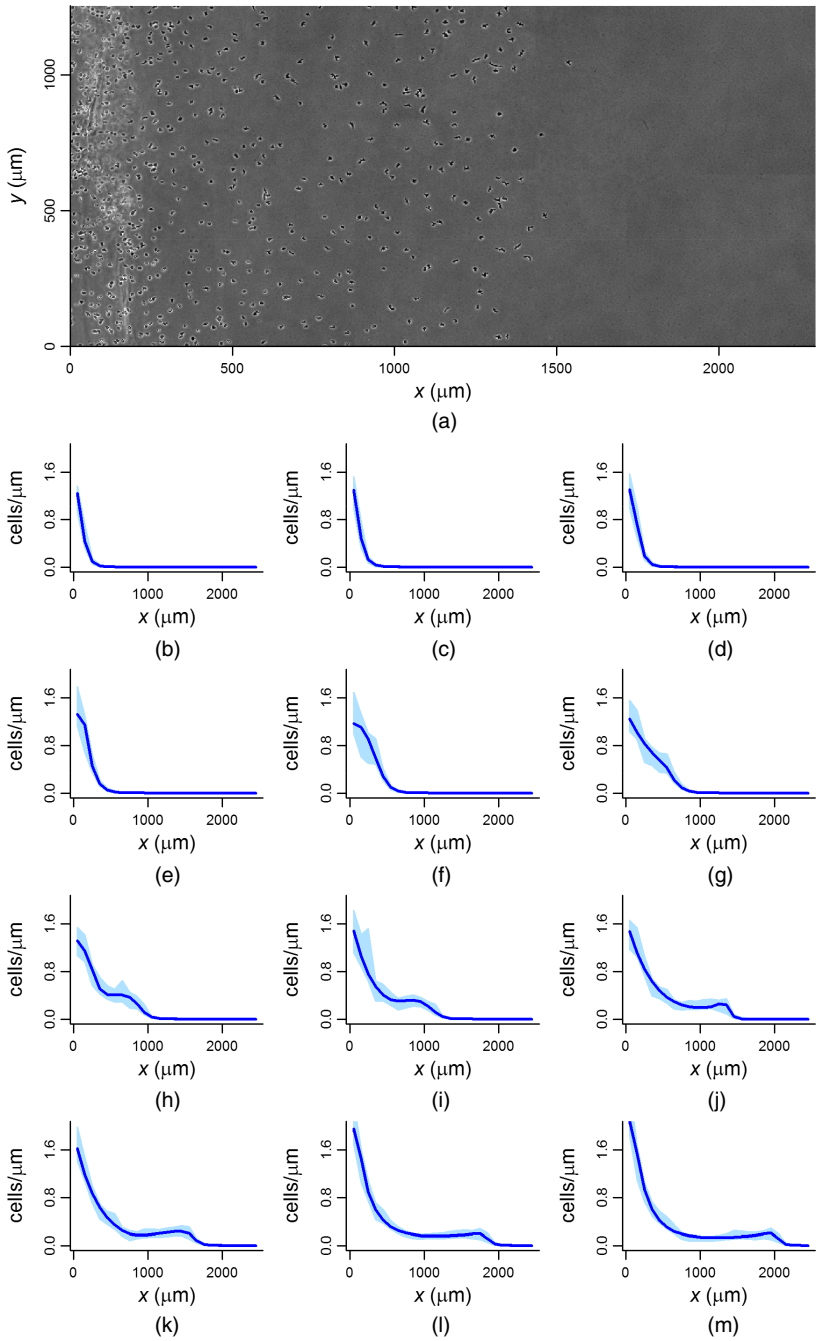

Fig. 1. (a) Example image from the *Dictyostelium discoideum* cell movement data set with 10 μM of folate in the gel (this image was obtained 4 h into the experiment (compare with (j)); the edge of the trough from which the cells originated is visible at the far left); (b)–(m) one-dimensional log-spline density estimates (Stone *et al.*, 1997) showing the cell distribution at half-hour intervals from the 10-μM folate data set (95-percentile intervals were obtained by non-parametric bootstrapping, using 10000 samples of the data) ((b) $t = 0$ h; (c) $t = 0.5$ h; (d) $t = 1$ h; (e) $t = 1.5$ h; (f) $t = 2$ h; (g) $t = 2.5$ h; (h) $t = 3$ h; (i) $t = 3.5$ h; (j) $t = 4$ h; (k) $t = 4.5$ h; (l) $t = 5$ h; (m) $t = 5.5$ h)

$$a(x,t) = \alpha(x,t)\,\frac{\partial A(x,t)}{\partial x} \tag{3}$$

where the parameter $\alpha \geqslant 0$ describes the responsiveness of the cells to the gradient (a larger α produces faster migration up the gradient). The attractant is depleted locally in proportion to the cell density and attractant concentration, according to the rate parameter $\gamma \geqslant 0$. Attractant may also diffuse through the medium, via a constant diffusion coefficient D_A, from areas of high to low concentration. Changes in attractant concentration through time can thus be described by

$$\frac{\partial A(x,t)}{\partial t} = -\gamma(t)\,C(x,t)\,A(x,t) + D_A\frac{\partial^2 A(x,t)}{\partial x^2}. \tag{4}$$

The efficiency of cell chemotaxis is known to depend on not just the strength of the chemotactic gradient, but also on the background concentration of the chemoattractant (Tweedy *et al.*, 2013). This is because cell receptors for an attractant become saturated when the attractant concentration is high, making it difficult for the cells to detect any underlying attractant gradient. What the cells are really responding to, therefore, is the gradient in saturation of their receptors. We incorporated this effect in our *receptor saturation model* as follows, by assuming that the interaction between folate and the cell folate receptors follows Michaelis–Menten kinetics:

$$a(x,t) = \alpha(x,t)\,\frac{\partial}{\partial x}\left\{\frac{A(x,t)}{A(x,t)+K_d}\right\}. \tag{5}$$

In this advection coefficient, the parameter K_d is the dissociation constant describing the folate concentration at which half the cells' folate receptors are occupied.

We developed the *interaction model* to investigate the effect of cells attracting or repelling one another through chemical interactions. In principle, this could be achieved by assuming that the cells release and respond to an additional chemoattractant or chemorepellant in their environment (Keller and Segel, 1970). However, this approach would require several additional model parameters, as well as an assumption about the initial concentration of this released attractant. A simpler alternative is to model cell interactions indirectly by assuming that cells seek to move up a spatial gradient in conspecific density (in addition to the gradient in attractant concentration $A(x,t)$) according to the advection coefficient:

$$a(x,t) = \alpha(x,t)\,\frac{\partial A(x,t)}{\partial x} + \frac{\eta(x,t)}{1+\lambda C(x,t)}\,\frac{\partial C(x,t)}{\partial x} \tag{6}$$

where η describes the strength of the response to the conspecific density gradient, describing attractive interactions when positive and repulsive interactions when negative. Since these cell interactions occur through chemical communication and require the cells to receive the signal via their receptors, we would expect a receptor saturation effect to occur at high cell densities when a large amount of the signal will be being released into the environment. This effect is implemented through the parameter $\lambda \geqslant 0$, which reduces the response to the cell density gradient when cell density increases. This parameter also helps to maintain model stability when $\eta > 0$ by reducing the chance that the cell density will blow up to unrealistic levels.

The cells may also affect one another's movement through direct physical contact. When densities become high, these contacts will increase in frequency, resulting in a blocking effect on movement. This effect can be added into the basic gradient model by altering the advection coefficient as follows to give our *overcrowding model* (see Hillen and Painter (2009) for a derivation):

$$a(x,t) = \left\{1 - \frac{C(x,t)}{C_{\max}}\right\}\left\{\alpha(x,t)\,\frac{\partial A(x,t)}{\partial x}\right\} \tag{7}$$

where $C_{\max}$ is the maximum cell density. Note that the rate of directional movement up the attractant gradient now declines as the cell density increases towards the maximum value.

The remaining four models in our nine-model candidate set have advection coefficients that include combinations of the three upgrades on the basic gradient model that we have described (i.e. receptor saturation, chemical cell interactions and overcrowding effects). The coefficients of these models are as follows:

(a) *receptor saturation and interaction model*,

$$a(x,t)=\alpha(x,t)\,\frac{\partial}{\partial x}\left\{\frac{A(x,t)}{A(x,t)+K_d}\right\}+\frac{\eta(x,t)}{1+\lambda C(x,t)}\,\frac{\partial C(x,t)}{\partial x}; \tag{8}$$

(b) *receptor saturation and overcrowding model*,

$$a(x,t)=\left\{1-\frac{C(x,t)}{C_{\max}}\right\}\left[\alpha(x,t)\,\frac{\partial}{\partial x}\left\{\frac{A(x,t)}{A(x,t)+K_d}\right\}\right]; \tag{9}$$

(c) *interaction and overcrowding model*,

$$a(x,t)=\left\{1-\frac{C(x,t)}{C_{\max}}\right\}\left\{\alpha(x,t)\,\frac{\partial A(x,t)}{\partial x}+\frac{\eta\,(x,t)}{1+\lambda C(x,t)}\,\frac{\partial C(x,t)}{\partial x}\right\}; \tag{10}$$

(d) *receptor saturation, interaction and overcrowding model*,

$$a(x,t)=\left\{1-\frac{C(x,t)}{C_{\max}}\right\}\left[\alpha(x,t)\,\frac{\partial}{\partial x}\left\{\frac{A(x,t)}{A(x,t)+K_d}\right\}+\frac{\eta(x,t)}{1+\lambda C(x,t)}\,\frac{\partial C(x,t)}{\partial x}\right]. \tag{11}$$

The cells may change their rates of movement and the rate at which they deplete folate as a result of changes in their state or environmental conditions. We have, therefore, allowed the model parameters describing the strength of these behaviours (α, D_C, γ and η) to vary in time. We expect spatial effects on the parameters to be limited because of the experimental set-up; the cells are moving under a gel, the structure and initial composition of which do not vary throughout the majority of the modelled region. However, the trough in which the cells begin the experiment is one major spatial feature in the cells' environment that could affect movement rates, as the cells will experience resistance as they move from the trough and under the gel (Laevsky and Knecht, 2001). The parameters directly controlling cell movement rates (α, D_C and η) are therefore allowed to vary in space in addition to time. The depletion rate of folate is expected to increase over time as the cells, induced by their exposure to folate, release increasingly more folate deaminase (the enzyme that is responsible for breaking down folate) into their environment (Bernstein *et al.*, 1981). However, there are no spatial features in the environment of our cells that could influence folate deaminase production (it will be unaffected by the presence of the trough for example). Hence, the folate depletion rate γ is allowed to vary in time, but not in space. Spatial and temporal dependence in η was implemented through the description

$$\eta(x,t)=E+F(x)+G(t) \tag{12}$$

where E is a constant, and $F(x)$ and $G(t)$ are polynomials, with zero intercepts, in space and time respectively. For α, D_C and γ, which are constrained to values 0 or greater, we exponentiated the right-hand side of equation (12); taking D_C as an example,

$$D_C(x,t)=\exp\{E+F(x)+G(t)\}. \tag{13}$$

Note that, for γ, the coefficients of $F(x)$ were set to 0. The degrees of the polynomials $F(x)$ and $G(t)$ were chosen through statistical model selection, as described in Section 5.

We formally adopt the hierarchical modelling framework that was proposed in Cressie and Wikle (2011), page 114, and specify probability distributions at three tiers of a basic hierarchy:

(a) data model, p(data | process,parameters);
(b) process model, p(process | parameters);
(c) parameter model, p(parameters).

The distribution at the top level corresponds to the observational noise model. We discuss the details in the on-line supplement C. As we show in that supplement, integrating over the process variables x leads to a convolution integral quantifying the discrepancy between the noisy measurements $\tilde{x}$ and the noise-free process variables x. However, owing to the high precision of confocal microscopy, quantified in supplementary Fig. C1, this error is negligible against the effect of the numerical discretization. The convolution integral thus effectively reduces to the convolution with a δ-function, which leaves the original function invariant. The probability distribution at the second tier is given by the solution of the PDEs (equation (1)), subject to a normalization operation, shown in equation (20) in Section 5.1. At the bottom level, we need to specify prior distributions on the parameters, as we describe in more detail in the following section.

3.2. Prior distribution

We could obtain literature values for two of our model parameters; the dissociation constant K_d (De Wit *et al.*, 1986) and the diffusion coefficient D_A (Kalimuthu and John, 2009; Ershad *et al.*, 2013) of folate. For D_A, where we had high confidence in the literature values due to their high level of consistency, we specified a rescaled beta prior, with mode positioned at the literature value and cut-offs positioned close to this value. For K_d, we specified a gamma prior with a mode of the literature value and scale chosen such that the probability fell to virtually 0 within an order of magnitude. These priors enforce the required positivity constraint.

Our knowledge of the experimental conditions allowed us to set sensible boundaries on the values of the parameters describing the initial distribution of folate, δ and ε (equation (14); see the on-line supplement D), so that the priors for these parameters could be described by using rescaled beta distributions.

For the remaining parameter priors, we used previous simulations based on an older independent data set from a different experiment, and we identified values of the parameters beyond which the cell distributions differed substantially from those observed. Priors were then defined on the basis of these extreme values as either normal distributions with mode 0 or exponential distributions, with scales chosen such that the probability of extreme values was close to 0. Full details of the priors applied in our study can be found in the on-line supplement D.

4. Numerical model solution

Numerical solution of our PDEs was carried out by using the method of lines (Schiesser and Griffiths, 2009). This involved discretizing the spatial region of interest of length l into a row of equal-sized boxes, so that changes in cell density and attractant concentration in these boxes through time could be described as a system of ordinary differential equations, which was solved numerically (see the on-line supplement E for details). For a given model and parameter set, we could thus obtain spatiotemporally varying functions describing cell density $C(x, t)$ and attractant concentration $A(x, t)$.

The initial cell density distribution $C(x, 0)$ was obtained for each data set from the cell locations at $t = 0$ by modelling the log-density function as a cubic spline. These splines were fitted by

maximum likelihood, with the number and location of knots being selected by using the Bayesian information criterion (Schwarz, 1978) via a stepwise knot addition and deletion procedure (Stone *et al.*, 1997). This log-spline density estimation technique was implemented in R (R Core Team, 2015) by using the logspline package (Kooperberg, 2015). The probability density function thus obtained was rescaled so that the integral of $C(x, 0)$ over the spatial region of interest was equal to the number of cell observations at $t = 0$. Data on the initial attractant distribution in the 10-μM folate data set were unavailable. However, a reasonable assumption is that this initial distribution follows a sigmoidal curve in one-dimensional space:

$$A(x, 0) = \frac{10}{1 + \exp\{-\delta(x - \varepsilon)\}}. \tag{14}$$

This curve gives low attractant concentrations to the left, where the attractant-free trough where the cells were introduced was, and rises smoothly as x increases to a maximum concentration of 10 μM (the homogeneous concentration of attractant in the gel before the introduction of any cells). The parameters δ and ε describe the steepness of the sigmoid and the point in space at which half the resources have been depleted respectively.

Over the experimental time periods there was considerable movement of cells into the region of interest via the left-hand boundary, which borders the trough where all the cells were originally (Fig. 1(a)). To replicate the movement of cells into the region in our models, the integral $\int_0^l C(x, t)\mathrm{d}x$, which describes the total number of cells at time t, must change through time at a rate that is informed by the data. This was achieved by first obtaining the time series

$$S = \{n_t : t \in (0, 0.5, 1.0, \ldots)\} \tag{15}$$

where n_t is the number of cells observed at time point t. We then fitted an interpolating spline $N(t)$ to S for each data set and thus obtained $N'(t)$, the rate at which the number of cells in the region of interest changed over time (on-line supplementary Fig. E1). Changes in cell numbers over time could potentially have arisen due to movements across the left-hand boundary from the trough area, movements across the two boundaries perpendicular to the trough (which were not included in our one-dimensional model) or cell replication. Since cells were as likely to move out of as into the region of interest across the boundaries perpendicular to the trough, such movements were a source of noise only and did not contribute to the systematic pattern of change in cell numbers. In addition, cell replication is expected to be a relatively minor contributor to increases in cell numbers over the time periods that were considered. Therefore, we assumed that the overall pattern of increasing cell numbers occurred because of movements into the region from the high-density trough area only, and we set the following boundary condition:

$$\mathrm{Flux}_C^{\mathrm{L}}(t) = N'(t) \tag{16}$$

where $\mathrm{Flux}_C^{\mathrm{L}}$ is the cell flux (i.e. the net movement of cells) across the left-hand boundary of the region. No cells reached the right-hand boundaries during the time periods that were considered, so we applied a zero-flux boundary condition:

$$\mathrm{Flux}_C^{\mathrm{R}}(t) = 0 \tag{17}$$

which prevented cell density from entering or leaving the region via this boundary. A summary of the cell movements in and out of the region of interest is provided in Fig. 2. For the attractant, we assumed that the fluxes across the region boundaries were equal to the fluxes across the nearest internal boundaries in the discretized spatial region, i.e.

$$\mathrm{Flux}_A^{\mathrm{L}}(t) = \mathrm{Flux}_A^{1,2}(t), \tag{18}$$

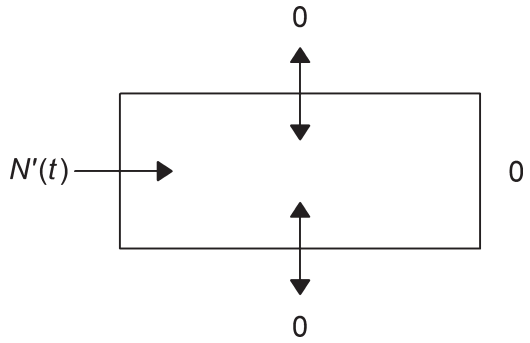

Fig. 2. Cell fluxes across the boundaries of the region of interest: the net movement across the left-hand boundary is given by $N'(t)$, movements in and out of the region balance out on the top and bottom boundaries and no movement occurs across the right-hand border

$$\mathrm{Flux}_A^{\mathrm{R}}(t) = \mathrm{Flux}_A^{B-1,B}(t) \tag{19}$$

where $\mathrm{Flux}_A^{1,2}$ is the cell flux between boxes 1 and 2, and B is the number of boxes making up the discretized region. We provide a description of how these boundary conditions were integrated into the numerical solution of the PDEs in the on-line supplement E.

5. Model inference

5.1. Likelihood calculation

Calculation of the likelihood of a given set of parameters $\boldsymbol{\theta}$ was achieved by using the cell density curve $C(x,t)$, which characterizes the cell distribution and is produced as an output of all the models investigated. The likelihood of $\boldsymbol{\theta}$ based on each observation $(y_1,\ldots,y_n)$ is given by

$$P(y_i|\boldsymbol{\theta}) = \frac{C(x_i,t_i)}{N(t_i)} \tag{20}$$

where x_i and t_i are the spatial location and time point that comprise y_i, i.e. $y_i = (x_i,t_i)$. Division by $N(t_i)$ is required to normalize the cell density and to convert it into a probability density, since, as a result of the cell initial and boundary conditions (equations (16) and (17)),

$$\int_0^l C(x,t)\,\mathrm{d}x = N(t). \tag{21}$$

The standard value of the total log-likelihood can be calculated by summation as

$$\log(L) = \sum_{i=1}^{n} \log\{P(y_i|\boldsymbol{\theta})\}. \tag{22}$$

Here, n is the total number of cells observed over the T time points, i.e.

$$n = \sum_{j=1}^{T} n_j \tag{23}$$

where n_j is the number of cells observed at time point $j \in \{0.5, 1.0, 1.5, \ldots\}$.

As cell numbers were greater for later time points because of movements into the region (272 observations at the first time point, compared with 757 at the final point), the standard log-likelihood may produce a fit that it is skewed towards these later time points. Therefore, we also consider the weighted log-likelihood

$$\log(\tilde{L}) = \frac{n}{T} \sum_{j=1}^{T} \left[\frac{1}{n_j} \sum_{i=1}^{n_j} \log\{P(y_i|\boldsymbol{\theta})\} \right] \tag{24}$$

where the contribution of each y_i is weighted on the basis of the number of observations at the associated time point. Multiplication by n/T brings the value back to the same scale as the standard log-likelihood. Weighted likelihoods have frequently been used to remove bias by downweighting observations that are believed to be of a lower quality (Hu and Zidek, 2002; Agostinelli and Greco, 2013). Here, we downweight observations not because they are of a lower quality, but because they provide us with less new information, given that we already have many other observations at the same time point.

5.2. Bayesian inference and model selection

We followed a Bayesian approach to inference and sampled parameters from the posterior distribution with Markov chain Monte Carlo (MCMC) sampling. The question is what kind of MCMC scheme to use. Standard random-walk Metropolis MCMC sampling turned out to be too slow in mixing. Advanced schemes, such as Hamiltonian Monte Carlo sampling, which require repeated likelihood computations along the proposal path, are computationally inefficient, because of the high computational costs of the numerical solution of the PDEs. A reasonable compromise is the delayed rejection adaptive Metropolis algorithm DRAM, proposed by Haario *et al*. (2006). This is an MCMC algorithm with a multivariate proposal distribution that is automatically adapted to allow for posterior correlations between the parameters and to identify the directions of principal change along the ridges in the posterior landscape. The acceptance rate is improved by the delayed rejection part of the algorithm where, instead of immediately advancing the chain following rejection of a parameter set, a second proposal is made that depends on both the current position of the chain and the rejected parameter set. Multiple additional proposals can be implemented if desired. We implemented DRAM by using the function `modMCMC` in the FME package (Soetaert and Petzoldt, 2010) in R (R Core Team, 2015), using one delayed rejection step, and updating the proposal distribution every 10 iterations.

The absence of any attractant in the experimental conditions that produced our 0-μM folate data set meant that we could immediately rule out all our models (Section 3) with the exception of the diffusion model (equations (1), (2) and (13)). We, therefore, use this data set to determine the appropriate degrees of the polynomials describing the dependences of the cell diffusion parameter D_C on space and time (equation (13)). A possible approach is to use reversible jump MCMC sampling (Green, 1995). However, convergence is typically slow, which is aggravated by the high computational costs of the numerical solution of the PDEs, and the parallel nature of the process. An alternative approach is the separate computation of marginal likelihoods; see for example Friel and Pettitt (2008). However, in combination with the numerical solution of the PDEs, the computational costs are unrealistically high. The method can in principle be parallelized, but in practice the parallel processing capacity is already used up by the parallel tempering scheme on which the method is based. An alternative approach, which is computationally less expensive, and promoted in Gelman *et al*. (2013), chapter 7, is the WAIC (Watanabe 2010), calculated as

$$\begin{aligned} \mathrm{WAIC} = &-2 \sum_{j=1}^{n} \log\left\{ \frac{1}{m} \sum_{i=1}^{m} P(y_j|\boldsymbol{\theta}_i) \right\} \\ &+2 \sum_{j=1}^{n} \left(\frac{1}{m} \sum_{i=1}^{m} \log\{P(y_j|\boldsymbol{\theta}_i)\}^2 - \left[\frac{1}{m} \sum_{i=1}^{m} \log\{P(y_j|\boldsymbol{\theta}_i)\} \right]^2 \right) \end{aligned} \tag{25}$$

where m is the number of sampled parameter sets, $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ are these parameter sets and $\mathbf{y} = (y_1, \ldots, y_n)$ are the observations. This score can be directly computed from the MCMC trajectory, and the computation is straightforward to parallelize, as the MCMC trajectories for different models can run on different processors simultaneously. We, therefore, fit versions of the diffusion model with polynomial degrees for the dependences of D_C on time and space ranging from 0 to 6 and select the best combination of polynomial degrees as that giving the lowest WAIC. Two chains were run from random parameters for each model variation, and we assessed within-chain convergence by using the Geweke diagnostic (Geweke, 1991) and between-chain convergence by using the Gelman–Rubin statistic (Gelman and Rubin, 1992).

For the 10-μM data set, we first took the degrees of the polynomials describing spatial and temporal dependences in D_C from the 0-μM folate data set and then carried out a local readjustment of these degrees by using the diffusion model applied to this new data set (see the on-line supplement G for details). We then ran MCMC simulations for the remaining eight candidate models using the 10-μM folate data. To keep the approach computationally feasible, we used the same polynomial degrees in space and time as were selected for D_C using the diffusion model for all four of our parameters with spatial and temporal dependences (α, η, γ and D_C) in the other models.

The advection terms entering all models other than the diffusion model are complex non-linear functions that model the processes of cell–cell interaction, cell–molecule interaction, membrane saturation etc. This has two consequences that affect MCMC convergence:

(a) the additional non-linear complexity changes the topology of the log-likelihood, leading to a higher degree of multimodality, and
(b) the system of coupled non-linear differential equations is stiff, leading to a substantial reduction in the numerical integration step size (for numerical stabilization).

The second aspect is particularly dramatic. We found that, by including the advection term, the numerical solution of the differential equations slows down by a whole order of magnitude as a mere consequence of the step size adjustment. Since the numerical solution of the differential equations is required in every step of the MCMC simulation, the effect on the overall run time is substantial: for the models other than the diffusion model, no indication of convergence was found despite a month of run time.

With the computational resources that were available to us, we could typically carry out 100 000 MCMC steps per week for our diffusion-only model, but only 10 000 MCMC steps per week for many of our more complex models with the non-linear advection terms included. To obtain a reasonable degree of convergence, quantified in terms of Rubin–Gelman potential scale reduction factors obtained from independent simulations started from hyperdispersed starting points, we would require far in excess of 100 000 MCMC steps for the models with the advection term included, which is computationally infeasible.

To deal with this problem, we adopted the following approximation. We started with repeated maximizations of the log-likelihood (more accurately: the log-unnormalized-posterior), to obtain a good approximation of the maximum *a posteriori* (MAP) parameter configuration. This exploits the fact that optimization is parallelizable, and that approximating the MAP configuration by the best local optimum from several independent initializations is common practice in complex systems science. We then started two independent MCMC simulations of a minimum 80 000 MCMC steps from the MAP configuration and checked for convergence on the basis of consistency of the WAIC scores obtained from two sections (the middle and end thirds of the MCMC chains, discarding the first third of steps as burn-in) from two independent MCMC runs (hence giving us four WAIC scores overall). In this way, we restrict the exploration of

Table 1. WAIC values for each model fitted to the 10-μM folate data set, using both the standard (equation (22)) and the weighted (equation (24)) likelihoods, L and $\tilde{L}$†

Model	*WAIC scores*	
	L	$\tilde{L}$
Diffusion	702.0 (0.1)	605.9 (0.09)
Gradient	4.3 (0.58)	4.5 (1.18)
Receptor saturation	13.5 (1.16)	15.6 (0.44)
Interaction	0 (0.55)‡	0 (1.52)‡
Overcrowding	3.5 (0.29)	3.4 (0.42)
Receptor saturation + interaction	12.0 (0.69)	6.7 (0.37)
Receptor saturation + overcrowding	12.4 (0.26)	11.5 (0.85)
Interaction + overcrowding	2.0 (1.39)	2.9 (0.73)
Receptor saturation + interaction + overcrowding	9.9 (0.9)	9.6 (1.55)

†The values for the diffusion model, which was the only model for which we achieved formal convergence of MCMC chains based on the Geweke and Gelman–Rubin diagnostics, were obtained by using equation (25), with the standard errors (in parentheses) being calculated as described in supplement I. The values for all other models were obtained as the means of the four WAIC values calculated from the mid- and end sections of the chains for those models (Fig. 4, supplementary Tables J1 and J2).
‡Best model.

the configuration space to the area around the MAP configuration. The justification of this approach is discussed in Section 7, and a test of the performance of the approach on simulated data is provided in the on-line supplement F. We repeated this procedure twice, using both the standard (equation (22)) and the weighted (equation (24)) likelihoods.

6. Results

WAIC values were obtained for fits of the diffusion model to the 0-μM folate data set with different combinations of polynomial orders for the dependences of the diffusion rate on space and time (equation (13)). We found that a polynomial degree of 2 in space and 4 in time was associated with the smallest WAIC values, both for the standard likelihood (equation (22); on-line supplementary Table G1) and the weighted likelihood (equation (24); supplementary Table G2). The cell distributions that were produced by this model show good agreement with those estimated directly from the data (Fig. 3). The patterns of change in cell diffusion in time and space that were predicted by this model are discussed in supplement H.

We fitted the diffusion model with a polynomial degree of 4 in time and 2 in space (as suggested by model selection on the 0-μM folate data set) to the 10-μM folate data set and then carried out a local readjustment of the polynomial degrees using this data set. This involved identifying polynomial coefficients where the posterior distribution was focused around zero (supplementary Fig. G3), and using this information as a guide to which polynomial degrees might be reduced to prevent unnecessary model complexity. We tried different adjustments of the polynomial degrees and selected the best degrees on the basis of the WAIC. This gave a degree of 3 in time for the standard likelihood and 2 for the weighted likelihood (supplementary Table G7). We maintained a polynomial degree of 2 in space for both the standard likelihood and the weighted likelihood, as suggested by supplementary Fig. G3.

WAIC values that were calculated from the mid- and end sections of the two chains for the eight models that include an advection component are closely grouped by model (Fig. 4), and

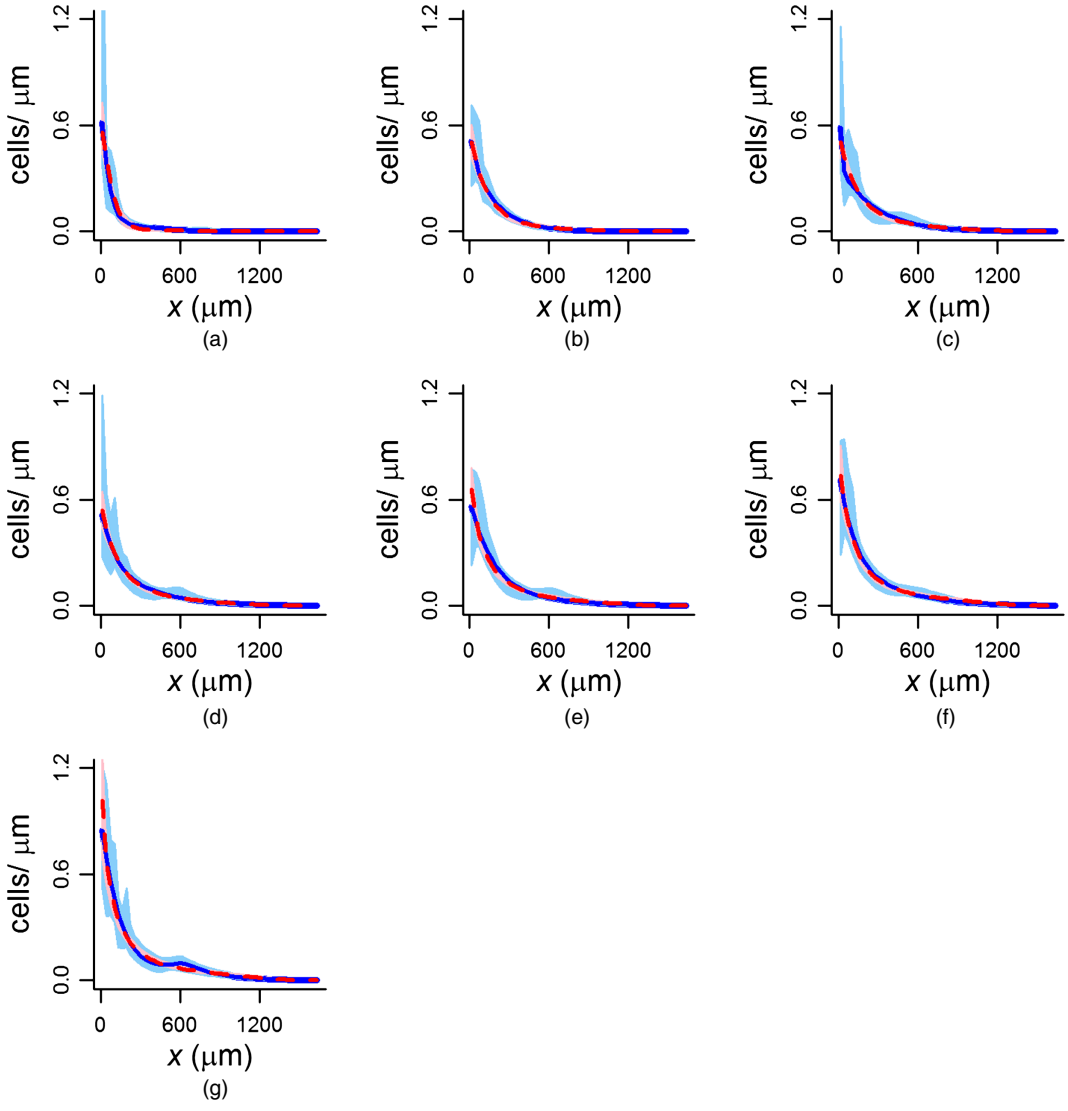

Fig. 3. Plots of the cell distributions (– – –) at half-hourly intervals simulated (using the posterior mean parameters) from the diffusion model fitted to the 0-μM folate data by using the standard likelihood (equation (22)), with polynomial degrees of 4 and 2 describing the temporal and spatial dependences of the diffusion coefficient respectively (direct density estimations (———) from the data, obtained by using log-spline density estimation (Stone *et al.*, 1997), are included for comparison; 95-percentile intervals for the density estimates (■) were obtained by non-parametric bootstrapping, using 10000 samples of the data; 95-percentile intervals for the model (■) were obtained from 500 samples from the posterior distribution): (a) $t = 0.5$ h; (b) $t = 1$ h; (c) $t = 1.5$ h; (d) $t = 2$ h; (e) $t = 2.5$ h; (f) $t = 3$ h; (g) $t = 3.5$ h

the ranking of the models based on these values is consistent across the standard and weighted likelihoods (Fig. 4 and Table 1). The diffusion model gives a much poorer WAIC value than the other models (Table 1), which all include an interaction of the cells with the chemoattractant (folate) in their environment, suggesting that this interaction is necessary for achieving a good fit to the data. For both the standard likelihood and the weighted likelihood, the interaction model produces the best mean WAIC value (Table 1), but there is a similar level of support for the model

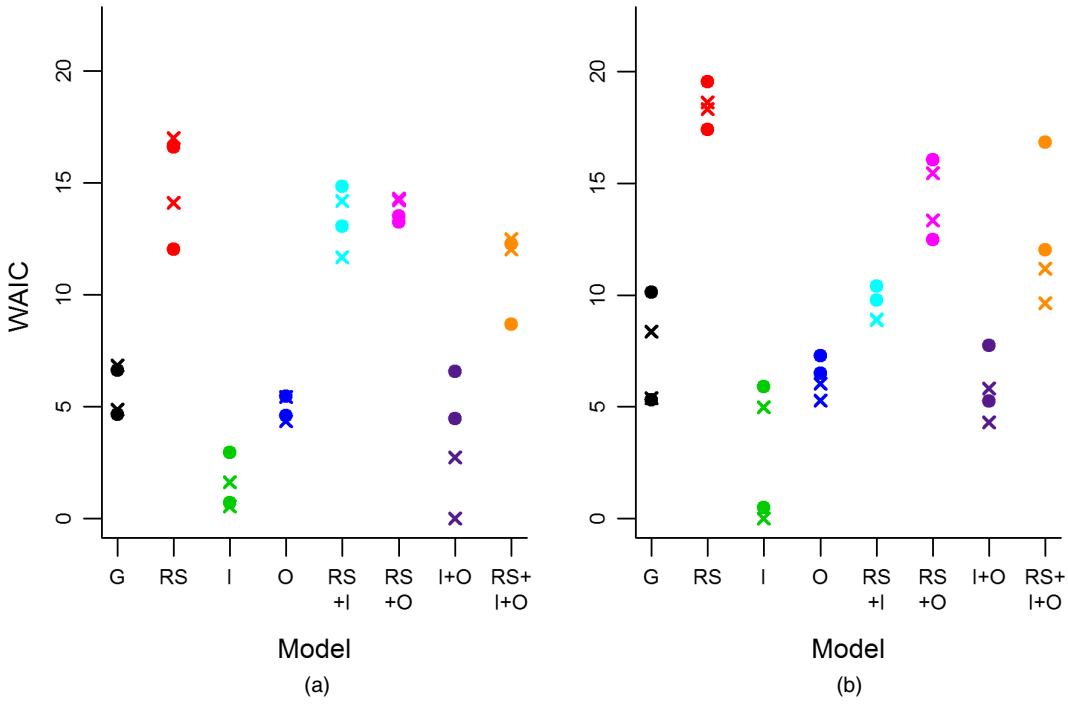

Fig. 4. Plots of the four WAIC values calculated for each of the models fitted to the 10-μM data set by using (a) the standard likelihood and (b) the weighted likelihood, L and $\tilde{L}$: for each model, we obtained two MCMC chains and calculated the WAIC (equation (25)) separately for the middle third (×) and the end third (●) of each chain (note that the minimum WAIC value has been subtracted from all values to aid comparison): G, gradient; RS, receptor saturation; I, interaction; O, overcrowding

that includes both interaction and overcrowding terms, as indicated by the standard errors of the mean WAIC values (Table 1), and the large degree of overlap between the four individual WAIC values for these models (Fig. 4). On examination of the parameters, we found that the estimated value of $C_{\max}$ (the maximum cell density), which implements the overcrowding effect that is described in equation (7)), was very large. A large value of $C_{\max}$ essentially causes the interaction and overcrowding model to revert to the interaction model, explaining the similarity in WAIC for these models. We, therefore, select the interaction model as the optimal model for explaining these data. In addition to concluding that the correction for overcrowding has, at most, a very small effect, we also find that the effect of receptor saturation does not improve model fit.

Model outputs from the interaction model show very good agreement with the 10-μM folate data (Fig. 5), successfully reproducing the steep cell front, which the simpler diffusion model fails to capture (supplementary Fig. K1). A discussion of the spatial and temporal dependences of the parameters for this model can be found in the on-line supplement H. A residual analysis finds no significant mismatch between our selected model and the data (see supplement L).

7. Discussion

We developed a detailed protocol for statistical inference in PDE models of cell migration and interaction. Formally, our mathematical description of the phenomenon resembles an advection–diffusion model, for which statistical inference has been reported in the literature before (Wikle

and Hooten, 2006; Cressie and Wikle, 2011). However, the key advance of our work is the considerably increased non-linearity in the 'advection' term, which describes a variety of processes that are related to the way that cells sense and interact with their environment. This leads to stiff PDEs, for which the numerical integration step size must be taken to be very small to stabilize the numerical solution, substantially increasing computational costs. Consequently, adequate adaptations are required to render statistical inference computationally viable.

We have adopted a Bayesian approach to inference, with a particular focus on model selection: given a set of hypotheses for the mechanisms driving cell migration, which are most consistent with the data? Model selection via Bayes factors, either directly estimated via parallel tempering (Friel and Pettitt, 2008), or indirectly by reversible jump MCMC sampling (Green, 1995), is computationally intractable because of the need to solve a stiff system of PDEs in every step of the Markov chain. Classical information criteria, however, such as the Akaike information criterion or the Bayesian information criterion, rely on asymptotics that are hardly met in practice, especially not for the high degree of non-linear complexity that is inherent in our model. As a compromise between numerical tractability and accuracy, we have adopted an approach based on the WAIC (Watanabe, 2010). This approach is similar to the deviance information criterion (Spiegelhalter *et al*., 2002) in spirit but has been shown to be more 'widely applicable' in the sense that it is *not* restricted to non-singular likelihood functions (as opposed to the deviance information criterion). The WAIC has been favourably reviewed in Gelman *et al*. (2013), chapter 7. A recent study that was carried out by one of the authors suggests that, for model selection in complex non-linear systems, the WAIC clearly outperforms the deviance information criterion and is on a par with Bayes factors (Aderhold *et al*., 2016).

We have found that the application of the procedure outlined to a diffusion model of the complexity of a 'standard' advection–diffusion model, e.g. as investigated in Wikle and Hooten (2006) and Cressie and Wikle (2011), is computationally tractable. However, when including the complex advection term, MCMC run times increase substantially as a consequence of the stiffness of the PDEs. This does not allow us to run MCMC simulations with a sufficient length to satisfy established convergence criteria. The method that we have proposed to deal with this difficulty is effectively a restriction of the configuration space. Rather than initializing independent MCMC simulations from starting points sampled from a hyperdispersed distribution, we started all MCMC simulations from the MAP parameters. We ran independent MCMC simulations over a minimum 80 000 iterations (the first third of which were discarded as burn-in) and computed the WAIC scores in a variety of ways: for different sections (middle *versus* end) of the same MCMC trajectory, for different MCMC trajectories, and for different objective functions (the standard *versus* the weighted log-likelihood). Our results show that the model selection results are consistent (Fig. 4). This suggests convergence in the actual WAIC scores offering confidence in our model selection results.

This method has the following justification.

(a) Approximating the posterior distribution by the area around the MAP configuration is akin to the Laplace approximation, which is widely applied to complex systems for which MCMC simulations are computationally too expensive (as evidenced by the large number of applications using the integrated nested Laplace approximation (Rue *et al*., 2009)). Our method is less restrictive than the Laplace approximation, in that it does not require a second-order truncation of the Taylor series expansion.

(b) Approximating the posterior distribution by a unimodal model distribution from a standard function family is also commonly done in variational inference, which is another alternative method for systems that are too complex for MCMC methods (e.g. Bishop

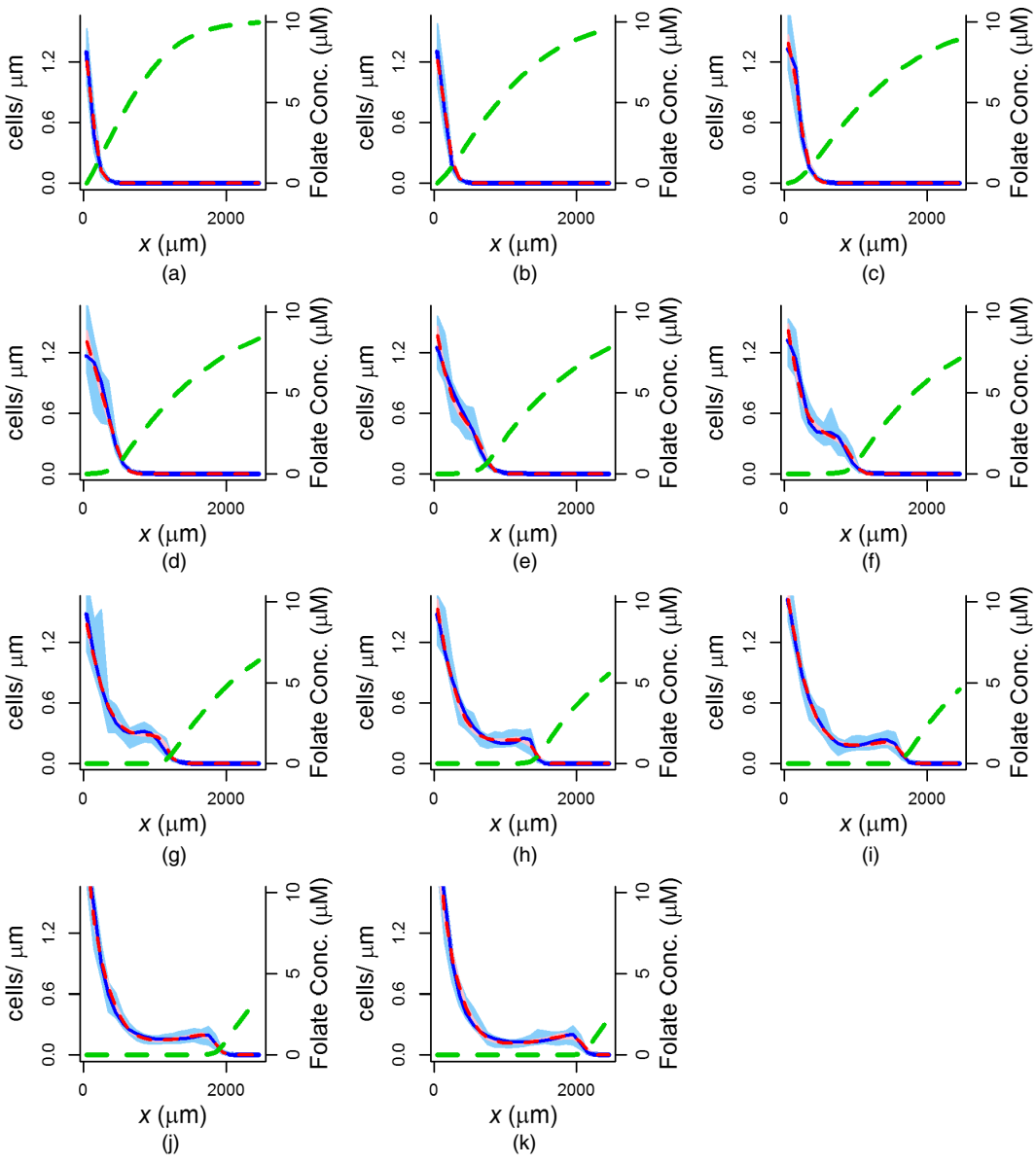

Fig. 5. Plots of the cell distributions at half-hourly intervals simulated from the interaction model fitted to the 10-μM folate data by using the standard likelihood (equation (22)) (——, density estimate; — —, cell model fit; — —, folate model fit) (we used the MAP parameter configuration of the model to produce the model fit lines; 95-percentile intervals for the model (▭) were obtained from 250 parameter sets sampled evenly from the last two-thirds of the two MCMC chains for this model; direct density estimations from the data, obtained by using log-spline density estimation, are included for comparison; 95-percentile intervals for the density estimates (▭) were obtained by non-parametric bootstrapping, using 10000 samples of the data): (a) $t = 0.5$ h; (b) $t = 1$ h; (c) $t = 1.5$ h; (d) $t = 2$ h; (e) $t = 2.5$ h; (f) $t = 3$ h; (g) $t = 3.5$ h; (h) $t = 4$ h; (i) $t = 4.5$ h; (j) $t = 5$ h; (k) $t = 5.5$ h

(2006)). Again, our approximation is less restrictive than variational inference, in that it does not restrict the approximation to any *a priori* chosen functional form.

In an empirical investigation using simulated data (see the on-line supplement F for details), we found that the level of accuracy and precision of our approach is the same as for model selection with Bayes factors calculated by using population MCMC methods (Girolami *et al*., 2010).

The only alternative approach that could achieve a degree of MCMC convergence that meets established convergence criteria is to resort to gradient matching (Xun *et al*., 2013). Here, the computational costs of the individual MCMC steps are substantially reduced by bypassing the need for a numerical solution of the PDEs. However, gradient matching is an approximate method, and the current state of the art incurs a potentially substantial loss in model accuracy (Macdonald *et al*., 2015).

Facing the choice between approximate modelling (gradient matching) and sound inference (standard MCMC convergence) *versus* accurate modelling (numerical integration) and approximate inference (MCMC sampling around the MAP configuration) we have opted for the latter alternative. This is in line with the frequently cited proposition by John W. Tukey (1915–2000) that

> 'the approximate answer to the right problem is worth a good deal more than an exact answer to the approximate problem'.

However, an interesting topic for future research is to put this proposition to the test and systematically to compare both paradigms empirically.

We have analysed the movement of *Dictyostelium* amoebae in an initially homogeneous distribution of the chemoattractant folate. This is a simple system with relatively few variables, which has been studied by using other approaches, but remains incompletely understood. We, and others, study it as an early step towards more complex models with greater unknowns, in particular tumour cell metastasis and tissue remodelling. By applying our proposed inference method and model selection using the WAIC to a set of nine candidate models, we have drawn three conclusions about the mechanisms that drive the *Dictyostelium* movements that were observed in our data. First, we find that a self-generated gradient in folate has a significant role in producing the observed movement patterns, as previously suggested by Tweedy *et al*. (2016). This self-generated gradient mechanism is responsible for the sharp, dense moving cell front that is characteristic of these data, and which simple diffusion models fail to replicate. Interest in self-generated gradients is growing rapidly, as studies have suggested that they may play an important role in embryonic development (Donà *et al*., 2013) and the spread of cancers (Muinonen-Martin *et al*., 2014). Many other examples of self-generated gradients probably remain to be discovered throughout biomedical science, as they have some unique properties, including range and robustness (Tweedy *et al*., 2016). Improved methods of detecting these gradients, such as the framework that we have described here, will therefore be important and desirable tools for future analyses. Our method provides a means of estimating how the form of the latent chemical gradient develops over time. This is generally not possible experimentally; measurement of the chemical gradient requires destruction of the gel under which the cells are moving and ends the experiment, making repeated measurements over time impossible (Tweedy *et al*., 2016). We find that the final shape of our estimated folate gradient is visually similar to that measured by Tweedy *et al*. (2016) at the end of a repeat of the same movement assay, suggesting that our model is performing well.

Our second finding is that including direct interactions between the cells, allowing them to attract or repel one another, provides an improvement in model performance, as indicated by

a reduction in the WAIC score. The *Dictyostelium* cells that were studied here were vegetative and therefore lack most of the complex cell–cell interactions of aggregating cells (Varnum and Soll, 1981; Loomis, 1982). However, vegetative cells still exhibit weaker interactions, including short-range cell–cell repulsion driven by autorepellents (Keating and Bonner, 1977; Kakebeeke *et al*., 1979). Additionally, lack of nutrients in the environment could cause the cells to starve progressively over the 5.5-h time period. During starvation, cells go through different phases of development, during which they produce cell surface molecules that affect movement by altering cell–cell interactions. Contact sites A, for example, are induced within hours of starvation (Eitle and Gerisch, 1977). Contact sites A mediate cell–cell adhesion and, although aggregation was not obvious in our data, low levels of contact sites A could still modify interactions between the cells. Changes in contact sites A and similar proteins could promote small repulsion and attraction effects, explaining why the interaction model was preferred. It is clear, however, that cell–cell interactions are not the primary driving mechanism of the observed movements; the improvement in WAIC that was obtained by including the interaction effect is smaller by a factor of 100 than that obtained by inclusion of the self-generated folate gradient (Table 1).

The third conclusion resulting from our analysis is that changes in cell behaviour in time and space can have substantial effects on the movement patterns that are observed (see the on-line supplement H for detailed descriptions and discussion of the inferred spatial and temporal dependences in the relevant model parameters). Local variations in the environment can alter the efficiency of cell movement over space. In our *Dictyostelium* system, for example, the edge of the trough within which the cells were seeded provides resistance to movement (Laevsky and Knecht, 2001). Additionally, many components underlying movement processes have been shown to vary in time. For example, the activities of both the folate receptor and the enzyme that is responsible for breaking down folate (folate deaminase) change over time in response to folate itself (Bernstein *et al*., 1981), and cells can rapidly vary the expression of multiple components of their motile and adhesive machineries. Despite this clear importance of spatiotemporal changes in cell behaviour, such changes are usually disregarded in assays and models for cell movement. Our method provides a new framework for analysing these dependences.

Despite its known influence on cell movement behaviour (Tweedy *et al*., 2013), we did not obtain an improvement in model performance on inclusion of the receptor saturation term. As we explain in the on-line supplement H, this surprising result is a consequence of the models without the receptor saturation term having enough flexibility to mimic the effect of receptor saturation through temporal and spatial variation in the basic gradient following mechanism.

In conclusion, we have presented a framework that allows effective inference and model comparison for complex PDE models, despite the serious computational costs that are incurred in solving these models numerically. This has allowed us to apply formal statistical inference in a field where it has previously been lacking; the modelling of cell movement. By carrying out model selection on an expansive set of candidate models, we could identify key mechanisms driving the movement of *Dictyostelium*, which is an organism with relatively simple cell movement behaviour, which has nonetheless been used to gain insights into aspects of human disease including immune defects and cancer (Carnell and Insall, 2011). Our models describe movement mechanisms, including self-generated chemoattractant gradients and cell–cell interactions, that are common to many cell types, making them widely applicable. By identifying the most likely mechanisms for movement in a particular system, the methods that were presented here could both guide future experimental work and suggest new medical interventions.

Acknowledgements

We thank Luke Tweedy for providing the data for analysis, and Diana Giurghita and Grant Hopcraft for helpful discussions of the concepts. Elaine Ferguson is funded by a University of Glasgow Lord Kelvin–Adam Smith scholarship.

References

Aderhold, A., Husmeier, D. and Grzegorczyk, M. (2016) Approximate Bayesian inference in semi-mechanistic models. *Statist. Comput.*, to be published.

Agostinelli, C. and Greco, L. (2013) A weighted strategy to handle likelihood uncertainty in Bayesian inference. *Computnl Statist.*, **28**, 319–339.

Bernstein, R. L., Rossier, C., Van Driel, R., Brunner, M. and Gerisch, G. (1981) Folate deaminase and cyclic AMP phosphodiesterase in Dictyostelium discoideum: their regulation by extracellular cyclic AMP and folic acid. *Cell Differ.*, **10**, 79–86.

Bishop, C. M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Carnell, M. J. and Insall, R. H. (2011) Actin on disease—studying the pathobiology of cell motility using Dictyostelium discoideum. *Semin. Cell Devlpmntl Biol.*, **22**, 82–88.

Coburn, L., Cerone, L., Torney, C., Couzin, I. D. and Neufeld, Z. (2013) Tactile interactions lead to coherent motion and enhanced chemotaxis of migrating cells. *Phys. Biol.*, **10**, article 046002.

Cressie, N. and Wikle, C. K. (2011) *Statistics for Spatio-temporal Data*. New York: Wiley.

Deisboeck, T. S. and Couzin, I. D. (2009) Collective behavior in cancer cell populations. *Bioessays*, **31**, 190–197.

De Wit, R. J., Bulgakov, R., Rinke de Wit, T. F. and Konijn, T. M. (1986) Developmental regulation of the pathways of folate-receptor-mediated stimulation of cAMP and cGMP synthesis in Dictyostelium discoideum. *Differentiation*, **32**, 192–199.

Donà, E., Barry, J. D., Valentin, G., Quirin, C., Khmelinskii, A., Kunze, A., Durdu, S., Newton, L. R., Fernandez-Minan, A., Huber, W., Knop, M. and Gilmour, D. (2013) Directional tissue migration through a self-generated chemokine gradient. *Nature*, **503**, 285–289.

Eitle, E. and Gerisch, G. (1977) Implication of developmentally regulated Concanavalin A binding proteins of Dictyostelium in cel adhesion and cyclic AMP regulation. *Cell Differ.*, **6**, 339–346.

Ershad, S., Dideban, K. and Faraji, F. (2013) Synthesis and application of polyaniline/multi walled carbon nanotube nanocomposite for electrochemical determination of folic acid. *Anal. Bioanal. Electrochem.*, **5**, 178–192.

Fleming, A. B. and Saltzman, W. M. (2002) Pharmacokinetics of the carmustine implant. *Clin. Pharmkinet.*, **41**, 403–419.

Fortin, D., Buono, P.-L., Fortin, A., Courbin, N., Gingras, C. T., Moorcroft, P. R., Courtois, R. and Dussault, C. (2013) Movement responses of caribou to human-induced habitat edges lead to their aggregation near anthropogenic features. *Am. Nat.*, **181**, 827–836.

Friel, N. and Pettitt, A. N. (2008) Marginal likelihood estimation via power posteriors. *J. R. Statist. Soc.* B, **70**, 589–607.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis*, 3rd edn. Boca Raton: CRC Press.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci.*, **7**, 457–472.

Geweke, J. (1991) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (eds J. Bernardo, J. Berger, A. Dawid and A. Smith). Oxford: Clarendon.

Girolami, M., Calderhead, B. and Vyshermirsky, V. (2010) System identification and model ranking: the Bayesian perspective. In *Learning and Inference in Computational Systems Biology*, 1st edn (eds N. D. Lawrence, M. Girolami, M. Rattray and G. Sanguinetti), pp. 201–230. Cambridge: MIT Press.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Haario, H., Laine, M., Mira, A. and Saksman, E. (2006) DRAM: efficient adaptive MCMC. *Statist. Comput.*, **16**, 339–354.

Hillen, T. and Painter, K. J. (2009) A user's guide to PDE models for chemotaxis. *J. Math. Biol.*, **58**, 183–217.

Hu, F. and Zidek, J. (2002) The weighted likelihood. *Can. J. Statist.*, **30**, 347–371.

Insall, R. H. (2010) Understanding eukaryotic chemotaxis: a pseudopod-centred view. *Nat. Rev. Mol. Cell Biol.*, **11**, 453–458.

Jaiswal, D. K. and Kumar, A. (2011) Analytical solutions of time and spatially dependent one-dimensional advection-diffusion equation. *Pollution*, **32**, 2078–2083.

Kakebeeke, P. I. J., De Wit, R. J. W., Kohtz, S. D. and Konijn, T. M. (1979) Negative chemotaxis in Dictyostelium and Polysphondylium. *Exptl Cell Res.*, **124**, 429–433.

Kalimuthu, P. and John, S. A. (2009) Selective electrochemical sensor for folic acid at physiological pH using ultrathin electropolymerized film of functionalized thiadiazole modified glassy carbon electrode. *Biosens. Bioelectron.*, **24**, 3575–3580.
Keating, M. T. and Bonner, J. T. (1977) Negative chemotaxis in cellular slime molds. *J. Bacteriol.*, **130**, 144–147.
Keller, E. F. and Segel, L. A. (1970) Initiation of slime mold aggregation viewed as an instability. *J. Theoret. Biol.*, **26**, 399–415.
Keller, E. F. and Segel, L. A. (1971) Model for chemotaxis. *J. Theoret. Biol.*, **30**, 225–234.
Kooperberg, C. (2015) logspline: logspline density optimality routines. *R Package Version 2.1.8*.
Laevsky, G. and Knecht, D. A. (2001) Under-agarose folate chemotaxis of Dictyostelium discoideum amoebae in permissive and mechanically inhibited conditions. *Biotechniques*, **31**, 1140–1149.
Loomis, W. F. (1982) *Development of Dictyostelium Discoideum*. New York: Academic Press.
Macdonald, B., Higham, C. and Husmeier, D. (2015) Controversy in mechanistic modelling with Gaussian processes. *Proc. 32nd Int. Conf. Mach. Learn.*, **37**, 1539–1547.
Majumdar, R., Sixt, M. and Parent, C. A. (2014) New paradigms in the establishment and maintenance of gradients during directed cell migration. *Curr. Opin. Cell Biol.*, **30**, 33–40.
McDonough, W. S., Johansson, A., Joffee, H., Giese, A. and Berens, M. E. (1999) Gap junction intercellular communication in gliomas is inversely related to cell motility. *Int. J. Dev. Neursci.*, **17**, 601–611.
Moorcroft, P. R., Lewis, M. A. and Crabtree, R. L. (2006) Mechanistic home range models capture spatial patterns and dynamics of coyote territories in Yellowstone. *Proc. R. Soc. Lond.* B, **273**, 1651–1659.
Muinonen-Martin, A. J., Susanto, O., Zhang, Q., Smethurst, E., Faller, W. J., Veltman, D. M., Kalna, G., Lindsay, C., Bennett, D. C., Sansom, O. J., Herd, R., Jones, R., Machesky, L. M., Wakelam, M. J. O., Knecht, D. A. and Insall, R. H. (2014) Melanoma cells break down LPA to establish local gradients that drive chemotactic dispersal. *PLOS Biol.*, **12**, article e1001966.
Neilson, M. P., Veltman, D. M., van Haastert, P. J. M., Webb, S. D., Mackenzie, J. A. and Insall, R. H. (2011) Chemotaxis: a feedback-based computational model robustly predicts multiple aspects of real cell behaviour. *PLOS Biol.*, **9**, article e1000618.
R Core Team (2015) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *J. R. Statist. Soc.* B, **71**, 319–392.
Scherber, C., Aranyosi, A. J., Kulemann, B., Thayer, S. P., Toner, M., Iliopoulos, O. and Irimia, D. (2012) Epithelial cell guidance by self-generated EGF gradients. *Integr. Biol.*, **4**, 259–269.
Schiesser, W. E. and Griffiths, G. W. (2009) *A Compendium of Partial Differential Equation Models: Method of Lines Analysis with Matlab*. Cambridge: Cambridge University Press.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Sibert, J. R., Hampton, J., Fournier, D. A. and Bills, P. J. (1999) An advection–diffusion–reaction model for the estimation of fish movement parameters from tagging data, with application to skipjack tuna (Katsuwonus pelamis). *Can. J. Fish. Aquat. Sci.*, **56**, 925–938.
Soetaert, K. and Herman, P. M. (2009) *A Practical Guide to Ecological Modelling: using R as a Simulation Platform*. New York: Springer.
Soetaert, K. and Petzoldt, T. (2010) Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME. *J. Statist. Softwr.*, **33**, 1–28.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Stone, C. J., Hansen, M. H., Kooperberg, C. and Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *Ann. Statist.*, **25**, 1371–1470.
Tweedy, L., Knecht, D. A., Mackay, G. M. and Insall, R. H. (2016) Self-generated chemoattractant gradients: attractant depletion extends the range and robustness of chemotaxis. *PLOS Biol.*, **14**, article e1002404.
Tweedy, L., Meier, B., Stephan, J., Heinrich, D. and Endres, R. G. (2013) Distinct cell shapes determine accurate chemotaxis. *Sci. Rep.*, **3**, article 2606.
Varnum, B. and Soll, D. R. (1981) Chemoresponsiveness to cAMP and folic acid during growth, development, and dedifferentiation in Dictyostelium discoideum. *Differentiation*, **18**, 151–160.
Venkiteswaran, G., Lewellis, S. W., Wang, J., Reynolds, E., Nicholson, C. and Knaut, H. (2013) Generation and dynamics of an endogenous, self-generated signaling gradient across a migrating tissue. *Cell*, **155**, 674–687.
Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, **11**, 3571–3594.
Wikle, C. K. and Hooten, M. B. (2006) Hierarchical Bayesian spatio-temporal models for population spread. In *Hierarchical Modelling for the Enviromental Sciences: Statistical Methods and Applications* (eds J. S. Clark and A. E. Gelfand), pp. 145–169. Oxford: Oxford University Press.
Wyckoff, J., Wang, W., Lin, E. Y., Wang, Y., Pixley, F., Stanley, E. R., Graf, T., Pollard, J. W., Segall, J. and Condeelis, J. (2004) A paracrine loop between tumor cells and macrophages is required for tumor cell migration in mammary tumors. *Cancer Res.*, **64**, 7022–7029.

Xun, X., Cao, J., Mallick, B., Maity, A. and Carroll, R. J. (2013) Parameter estimation of partial differential equation models. *J. Am. Statist. Ass.*, **108**, 37–41.
Zlatev, Z., Berkowicz, R. and Prahm, L. P. (1984) Implementation of a variable stepsize variable formula method in the time-integration part of a code for treatment of long-range transport of air pollutants. *J. Computnl Phys.*, **55**, 278–301.
Zoppou, C. and Knight, J. H. (1997) Analytical solutions for advection-diffusion equations with spatially variable coefficients. *J. Hydraul. Engng*, **123**, 144–148.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary information for "Statistical inference of the mechanisms driving collective cell movement"'.